CHAPITRE 3: ÉCHANTILLONNAGE ET ESTIMATION

1. Echantillonnage

1.1. **Notion d'échantillonnage.**

Définition 1.1. On considère une population Ω de taille N. On appelle **échantillon** un sous-ensemble de cette population. Un échantillon de taille n est donc une liste de n individus $(\omega_i, \omega_2, ..., \omega_n)$ extraits de lapopulation mère.

Exemple 1.1. *On considère une population constituée de 5 étudiants et on s'intéresse au temps hebdomadaire consacré par chaque étudiant à l'étude des statistiques.* $\Omega =$

Table 1. Tab1

Etudiant	*Temps d'étude (h)*
A	*7*
B	*3*
C	*6*
D	*10*
E	*4*

A, B, C, D, E *et* $N = 5$.

Définition 1.2. On appelle **échantillonnage** le prélèvement d'échantillons. Le rapport t de l'effectif n de l'échantillon sur l'effectif N de la population dans laquelle il a été prélevé, est appelé **taux d'échantillonnage** ou **fraction de sondage** i.e. $t = \dfrac{n}{N}$

Exemple 1.2. *On prélève des échantillons de taille* 2 *on* $t = \dfrac{2}{5}$ *(voir l'exemple 1.1)*

Définition 1.3. On appelle **échantillonnage aléatoire** un prélèvement de n individus dans une population mère tel que toutes les combinaisons possible de n individus aient la même probabilités d'être prélevés.

Il existe d'autres formes d'échantillonnage, on ne s'intéressera néanmoins qu'à des échantillonnage aléatoires.

Remarque: On cherche à décrire un caractère C qualitatif ou quantitatif présent dans une population Ω à travers l'étude des résultats obtenus sur un échantillon de taille n.

Exemple 1.3.
(1) *Étant donnée une population, on peut s'intéresser aux caractère quantitatif tels que le poids, la taille, ...etc.*
(2) *Étant donnée une population, on peut s'intéresser aux caractère qualitatif tels que la couleur des yeux, la couleur des cheveux, ...etc*
(3) *Le caractère étudié dans l'exemple initial est le temps hebdomadaire consacré à l'étude des statistiques.*

Définition 1.4. Soit C un caractère quantitatif défini sur une population mère Ω. C est la réalisation d'une variable aléatoire X définie sur Ω:

$$X : \Omega \longrightarrow \Re$$

$$\omega_i \longrightarrow X(\omega_i) = x_i$$

On appelle $n-$**échantillon de valeur de** X la liste des valeurs $(x_1, x_2, ..., x_n)$ observées prises par X sur un échantillons $(\omega_1, ..., \omega_n)$ de la population Ω. Les coordonnées peuvent être considérées comme les valeurs des réalisations d'un vecteur de variables aléatoires $(X_1, ..., X_n)$ appelé $n-$**échantillon de X** où les X_i sont de même loi, indépendantes.

Définition 1.5. On appelle **statistique** toute variable aléatoire qui s'écrit à l'aide des variables aléatoires $X_1, ..., X_n$.

Exemple 1.4. *X_i, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ sont des statistiques.*

Si on extrait plusieurs échantillons de taille n fixée, les résultats que l'on va pouvoir déduire sont variables car ils dépendent de l'échantillon considéré. On parle de **fluctuations d'échantillonnage**. Comment dans ce cas tirer des conclusions valables sur la population mère? On va pour cela étudier les loi de probabilité qui régissent ces fluctuations.

1.2. **Distributions d'échantillonnage.**

1.2.1. *Moyenne d'échantillon - Variance d'échantillon.*

Définition 1.6. On considère une population Ω dont les éléments possèdent un caractère quantitatif C qui est la réalisation d'une variable aléatoire X qui suit une loi de probabilité d'espérance μ et d'écart-type σ. On suppose que la famille est de taille infinie ou que l'échantillonnage se fait avec remise.
On prélève un échantillon $(X_1, ..., X_n)$ de X de valeurs $(x_1, ..., x_n)$. La moyenne $\overline{x}$ de l'échantillon est donnée par

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

il s'agit de la valeur prise par la variable aléatoire

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

La variable aléatoire $\overline{X}$ appelé **moyenne d'échantillon**. De la même manière la variance v de l'échantillon $(x_1, ..., x_n)$ est donnée par

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Il s'agit de la valeur prise par la variable aléatoire

$$v = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

On définit la variable aléatoire S^2, appelé **variance d'échantillon**, par

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}V.$$

Paramètre descriptifs de la distribution. Proposition:

(1) Quelle que soit la loi de X, on a

$$E(\overline{X}) = \mu$$

$$E(V) = \frac{n-1}{n}\sigma^2$$

et $Var(\overline{X}) = \dfrac{\sigma^2}{n},\ E(S^2) = \sigma^2.$

(2) Si $X \sim N(\mu, \sigma)$, on a

(i) si σ est connu: $\dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0,1)$,

(ii) si σ n'est pas connu (inconnu): $\dfrac{\overline{X} - \mu}{\sqrt{\dfrac{V}{n}}} \sim T_{n-1}$,

(iii) $\dfrac{nV}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

1.3. **Proportion d'échantillon.** Il arrive que le caractère à estimer ne soit pas quantitatif mais qualitatif. Dans ce cas, on recherche la proportion p des individus présentant ce caractère.

La proportion p sera estimée à l'aide des résultats obtenus sur un $n-$échantillon.

Définition 1.7. La proportion f obtenue dans un $n-$échantillon est la valeur observée d'une variable aléatoire F, fréquence d'apparition de ce caractère dans un échantillon de taille n, appelée **proportion d'échantillon** ou **fréquence statistique**.

On peut écrire

$$F = \frac{K}{n}$$

où K est la variable aléatoire qui compt le nombre d'apparitions du caractère considéré dans un échantillon de taille n.
Par définition, $K \sim B(n,p)$, soit

$$E(K) = np,\ Var(K) = npq.$$

t.q $q = 1 - p$.

Proposition: $E(F) = p$ et $Var(F) = \dfrac{pq}{n}$.
Remarque: Pour $n \geq 30,\ np \geq 15$ et $nq \geq 15$ on peut approcher F par une loi normale $N(p, \sqrt{\dfrac{pq}{n}}$.

2. Estimation

2.1. **Les estimateurs.** Estimer un paramètre c'est en recherche une valeur approchée à partir des résultats obtenus sur un échantillon.

Exemple 2.1. *Estimer la taille moyenne d'une population à partir de la moyenne empirique obtenue sur un échantillon de cette population.*

Définition 2.1. un **estimateur** $\widehat{\theta}$ du paramètre inconnu θ est une fonction qui fait correspondre à une suite d'observations une valeur approchée $\widehat{\theta}$ de θ, appelée estimation

$$\widehat{\theta} : (x_1, ..., x_n) \longmapsto \widehat{\theta} = f(x_1, ..., x_n).$$

Un estimateur $\widehat{\theta}$ est donc une variable aléatoire, on peut en calculer son espérance $E(\widehat{\theta})$ et sa variance $Var(\widehat{\theta})$. Ces quantité vont permettre de déterminer la qualité d'un estimateur du paramètre θ à estimer.
Un paramètre peut en effet avoir plusieurs estimateurs. Dans le cas de la taille moyenne d'une population, on peut choisir la moyenne arithmétique, la médiane, etc.

Définition 2.2. On dit que $\widehat{\theta}$ est un **estimateur sans biais** si la moyenne de sa distribution d'échantillonnage est égale à la valeur θ du paramètre à estimer

$$E(\widehat{\theta}) = \theta.$$

Sinon, on parle d'**estimateur biais**. Pour comparer les estimateurs biaisés, on introduit la quantité suivante:
On appelle **biais** d'un estimateur $\widehat{\theta}$ la quantité

$$Biais(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

Remarque: L'absence de biais n'est pas suffisante pour assurer de l'efficacité d'un estimateur. Le paramètre θ peut d'ailleurs présenter plusieurs estimateurs sans biais. Dans ce cas, c'est la variance des estimateurs qui permet de les comparer. Si cette variance est élevée, l'estimateur peut prendre des valeurs très éloignées de la valeur effective du paramètre θ.

Définition 2.3. On dit qu'un estimateur sans biais est **efficace** si sa variance est la plus petite parmi les variance des estimateurs sans biais. Si $\widehat{\theta}_1$ est un estimateur de θ, on dit que $\widehat{\theta}_1$ est efficace si pour tout estimateur sans biais $\widehat{\theta}_2/$

$$E(\widehat{\theta}_1) = E(\widehat{\theta}_2)$$

et

$$Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2)$$

Définition 2.4. Un estimateur $\widehat{\theta}$ est **convergent** si sa distribution tend à se concentrer autour de la valeur θ à estimer, en d'autre termes si sa variance tend vers zéro lorsque la taille de l'échantillon augmente:

$$\lim_{n \longrightarrow +\infty} Var(\widehat{\theta}) = 0.$$

2.1.1. *estimateurs usuels. (A) Cas d'un caractère quantitatif*

Soit X une variable aléatoire de moyenne μ et d'écart-type σ définie sur une population mère Ω. Soit $(X_1, ..., X_n)$ un n-échantillon de X.

Propriétés: Ona les résultats suivants:

(1) $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais et convergent de μ (i.e. $E(\overline{X}) = \mu$).
(2) $V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ est estimateur biais de la variance σ^2.
(3) $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n}{n-1}V$ est un estimateur sans biais et convergent de la variance σ^2.

(B) Cas d'un caractère qualitatif

On considère un caractère qualitatif d'une population dont on cherche à estimer la proportion p.

Propriété: La proportion d'échantillon F est un estimateur sans biais et convergent de la proportion p.

2.2. **Intervalles de confiance.** Plutôt que de déterminer une valeur approchée d'un paramètre θ obtenue à l'aide d'un estimateur $\overline{\theta}$, on va recherche un intervalle dans lequel on sait avec une probabilité satisfaisante que la valeur de θ s'y trouve.

Définition 2.5. Soit X une variable aléatoire dont la loi dépend d'un paramètre θ. Les **intervalles de confiance de risque** α pour le paramètre θ, issue des différents $n-$échantillons $(x_1, ..., x_n)$, sont les intervalles $[a(x_1, ..., x_n); b(x_1, ..., b_n)]$ tels qu'une proportion α de ces intervalles contiennent θ.

Remarque:

(1) La quantité $1-\alpha$ est appelée **niveau de confiance** de l'intervalle $[a,b]$:

$$P(a \leq \widehat{\theta} \leq b) = 1-\alpha.$$

(2) Dans la pratique, on ne dispose bien souvent que d'un seul échantillon qui fournit un intervalle de confiance $[a,b]$.

(3) Le paramètre à estimer est souvent l'espérance ou la variance dans le cas d'un caractère quantitatif, la proportion dans le cas d'un caractère qualitatif.

Dans la suite à s'attachera à recherche des intervalle de confiance $[a,b]$ symétriques, c'est à dire tels que:

$$P(\widehat{\theta} < a) = \frac{\alpha}{2}$$

et

$$P(\widehat{\theta} > b) = \frac{\alpha}{2}.$$

On détermine ensuite les variables aléatoires A_n et B_n en fonction de $\widehat{\theta}$ telles que:

$$P(A_n \leq \theta \leq B_n) = 1-\alpha.$$

Un intervalle de confiance $[a,b]$ de risque pour θ, issu d'un $n-$échantillon $(x_1,...,x_n)$ de valeurs de X, s'obtient alors en calculant:

$$a = A_n(x_1,...,x_n),\ b = B_n(x_1,...,x_n).$$

2.2.1. *Intervalle de confiance pour une moyenne.* On se place dans le cas où X suit une loi normale de paramètres μ et σ ou bien dans le cas où l'on ne connait pas forcément la loi de X mais pour laquelle on dispose d'un échantillon de taille $n > 30$. Dans le premier cas $\widehat{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, dans le second cas $\overline{X}$ suit approximativement cette même loi.

On considère un $n-$échantillon $(x_1,...,x_n)$ de valeurs de X. On note

$$m = \frac{x_1 + ... + x_n}{n}$$

et

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2.$$

(A) Cas σ connu

On sait que $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, soit encore

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(1, 0)$$

Donc

$$P(-t_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < -t_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

alors

$$P(\overline{X} - t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$$

Donc

$$Ic = [m - t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; m + t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

(B) Cas σ inconnu

$$Ic = [m - t_{1-\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}; m + t_{1-\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}]$$

où $t_{1-\frac{\alpha}{2},n-1}$ est le quantité d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $n-1$ degrés de liberté.

Remarque: Si $n > 30$, $t_{1-\frac{\alpha}{2},n-1} = t_{1-\frac{\alpha}{2}}$.

2.2.2. *Intervalle de confiance pour une variance.* On se place dans le cas le cas où X suit une loi normale de paramètres μ et σ.

(A) Cas μ connu

$$Ic = [\frac{nv}{\chi^2_{1-\frac{\alpha}{2}}(n)}; \frac{nv}{\chi^2_{\frac{\alpha}{2}}(n)}]$$

où $\chi^2_{1-\frac{\alpha}{2}}(n$ et $\chi^2_{\frac{\alpha}{2}}(n)$ sont les quantiles d'ordre $1 - \frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la loi de chi-deux à n degrés de liberté et

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

(B) Cas μ inconnu

$$Ic = [\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}; \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}]$$

où $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ et $\chi^2_{\frac{\alpha}{2}}(n-1)$ sont les quantiles d'ordre $1-\frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la loi de chi-deux à $n-1$ degrés de liberté.

remarque: si $n > 30, \chi^2_\alpha(n-1) \approx \frac{1}{2}(t_\alpha + \sqrt{2n-3})^2$, si bien que l'on choisit:

$$Ic = [\frac{2(n-1)s^2}{(t_{1-\frac{\alpha}{2}} + \sqrt{2n-3})^2}; \frac{2(n-1)s^2}{(t_{\frac{\alpha}{2}} + \sqrt{2n-3}})^2]$$

d'autre part, la symétrie de la loi normale centrée réduite assure que $t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$.

2.3. **Intervalle de confiance pour une proportion.** On a vu dans la partie précédente que la proportion d'échantillon F peut être approchée par une loi normale $N(p, \sqrt{\frac{pq}{n}})$ t.q. $q = 1-p$.
On en déduit:

$$Ic = [f - t_{1-\frac{\alpha}{2}}\sqrt{\frac{f(f-1)}{n}}; f + t_{1-\frac{\alpha}{2}}\sqrt{\frac{f(f-1)}{n}}]$$

où f est la proportion de l'échantillon analysé.